



A

By Express Mail # EM262493442US • February 9, 1998

Assistant Commissioner for Patents  
Washington, DC 20231

Attorney Docket #: 4577-50

Sir:

Transmitted herewith for filing is the utility patent application of:

Inventor(s): Jung Chul LEE, Min Soo HAHN, Hang Seop LEE

For: A Text-To-Speech Conversion System For Interlocking With Multimedia And A  
Method For Organizing Input Data Of The Same

Enclosed are:

1. Specification (19 p.), Claims 1 to 12 (5 p.) & Abstract (2 p.)
2. Executed Declaration and Power of Attorney (3 p.)
3. 2 sheet(s) of drawing(s) (Figs. 1 to 3)
4. Assignment of the invention to Electronics and Telecommunications Research Institute
5. Recordation Cover Sheet (PTO-1595)
6. Check for \$40.00 for Assignment Recording Fee
7. Preliminary Amendment
8. Verified Statement to Establish Small Entity Status
9. Certified copy of priority document No. 97-17615
10. Letter Transmitting Priority document
11. Information Disclosure Statement
12. PTO Form 1449 with copies (1 doc.) of cited reference

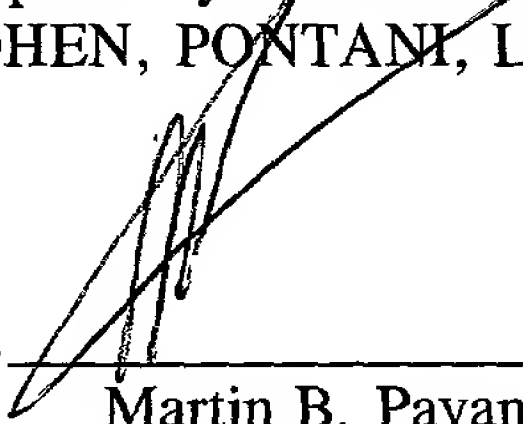
The filing fee has been calculated as shown below:

FOR:	Col. 1	Col. 2	SMALL ENTITY	OTHER THAN SMALL ENTITY
	# FILED	# EXTRA		
BASIC FEE			\$395	\$790
TOTAL CLAIMS	12 - 20 =	0	x 11 = \$	x 22 = \$
INDEPENDENT CLAIMS	2 - 3 =	0	x 41 = \$	x 82 = \$
<input type="checkbox"/> MULTIPLE DEPENDENCY			+ 135 = \$	+ 270 = \$
* If the difference in Col. 1 is less than zero, enter "0" in Col. 2			TOTAL: \$ 395	TOTAL: \$

- ☐ Please charge my Deposit Account No. 03-2412 in the amount of \$\_. A duplicate copy of this sheet is enclosed.
- ☒ A check in the amount of \$ 395 to cover the filing fee is enclosed.
- ☒ The Commissioner is hereby authorized to charge payment of the following fees associated with this application or credit any overpayment to Deposit Acct. No. 03-2412.
- ☒ Any additional filing fees required under 37 CFR 1.16.
- ☒ Any patent appl. processing fees under 37 CFR 1.17
- ☒ The issue fee set in 37 CFR 1.18 at 3 months from mailing of the Notice of Allowance, pursuant to 37 CFR 1.311 (b) provided the fee has not already been paid by check.
- ☒ Any filing fees under 37 CFR 1.16 for presentation of extra claims.
- ☒ Priority is claimed for this invention and application, corresponding applications having been filed in Korea on May 8, 1997, Application No. 97-17615.

Respectfully submitted,  
COHEN, PONTANI, LIEBERMAN & PAVANE

551 Fifth Avenue, Suite 1210  
New York, New York 10176  
Tel: (212) 687-2770  
Fax: (212) 972-5487

By:   
Martin B. Pavane  
Reg. No. 28,337

Dated: February 9, 1998

A TEXT-TO-SPEECH CONVERSION SYSTEM FOR INTERLOCKING WITH  
MULTIMEDIA AND A METHOD FOR ORGANIZING INPUT DATA OF THE SAME

BACKGROUND OF THE INVENTION

Field of the Invention

The present invention relates to a text-to-speech conversion system (hereinafter, referred to as TTS) for interlocking with multimedia and a method for organizing input data of the same, and more particularly to a text-to-speech conversion system (TTS) for interlocking with multimedia and a method for organizing input data of the same for enhancing the natural of synthesized speech and accomplishing synchronization between multimedia and TTS by defining additional prosody information, the information required to interlock TTS with multimedia, and interface between these information and TTS for use in the production of the synthesized speech.

Description of the Related Art

Generally, the function of the speech synthesizer is to provide different forms of information for a man using a computer. To this end, the speech synthesizer should serve the user with synthesized speech with high quality from a given text. In addition, for the interlock with database produced in multimedia environment such as moving picture or animation, or

a variety of media provided from a counterpart of conversion, the speech synthesizer should produce the synthesized speech to be synchronized with these media. Particularly, the synchronization of TTS with multimedia is essential to provide the user with service with high quality.

As shown in Fig. 1, typically, a conventional TTS goes through the process consisting of 3 steps as follows until the synthesized speech is produced from an inputted text.

In a first step, a language processor 1 converts the text into a series of phoneme, presumes prosody information and symbolizes this information. Symbol of prosody information is presumed from a boundary of the phrase and paragraph, a location of accent in word, a sentence pattern, and so on using the analysis result of syntax.

In a second step, a prosody processor 2 calculates a value of prosody control parameter from the symbolized prosody information using a rule and a table. Prosody control parameter includes duration of phoneme, pitch contour, energy contour, and pause interval information.

In a third step, a signal processor 3 produces a synthesized speech using a synthesis unit database 4 and the prosody control parameter. In other words, this means that the conventional TTS should presume the information associated with the natural and speech rate in the language processor 1 and the prosody processor

2 only by the inputted text.

Further, the conventional TTS has simple function to output data inputted by the unit of sentence as the synthesized speech. Accordingly, in order to output sentences stored in a file or sentences inputted through a communication network as the synthesized speech in succession, a main control program which reads sentences from the inputted data and transmits them to an input of TTS is required. Such a main control program includes a method to separate the text from the inputted data and then output the synthesized speech once from the beginning to the end, a method to produce the synthesized speech in interlock with a text editor, a method to look up the sentences by use of a graphic interface and produce the synthesized speech, and so on, but the object to which these methods are applicable is restricted to the text.

At present, studies on TTS have considerably advanced for the vernacular language in different countries and a commercial use has been accomplished in some countries. However, this is in situation of the only use for the syntheses of speech from the inputted text. In addition, by a prior organization, since it is impossible to presume from only the text the information required when moving picture is to be dubbed by use of TTS or when the natural interlock between the synthesized speech and multimedia such as animation is to be implemented, there is no method to realize these functions. Furthermore, there is also no result of the studies on use of additional data for

enhancement of the natural in the synthesized speech and organization of these data.

#### SUMMARY OF THE INVENTION

Therefore, it is an object of the present invention to provide a text-to-speech conversion system (TTS) for interlocking with multimedia and a method for organizing input data of the same for enhancing the natural of synthesized speech and accomplishing synchronization of multimedia with TTS by defining additional prosody information, the information required to interlock TTS with multimedia, and interface between these information and TTS for use in the production of the synthesized speech.

In order to accomplish the above object, a TTS for interlocking with multimedia according to the present invention comprises a multimedia information input unit for organizing text, prosody, the information on synchronization with moving picture, lip-shape, and the information such as individual property; a data distributor by each media for distributing the information of the multimedia information input unit into the information by each media; a language processor for converting the text distributed by the data distributor by each media into phoneme stream, presuming prosody information and symbolizing the information; a prosody processor for calculating a value of prosody control parameter from the symbolized prosody information using a rule and a table; a synchronization adjustor for

adjusting the duration of the phoneme using the synchronization information distributed by the data distributor by each media; a signal processor for producing a synthesized speech using the prosody control parameter and data in a synthesis unit database; and a picture output apparatus for outputting the picture information distributed by the data distributor by each media onto a screen.

In order to accomplish the above object, a method for organizing input data of a text-to-speech conversion system (TTS) for interlocking with multimedia comprises the steps of: classifying multimedia input information organized for enhancing the natural of synthesized speech and implementing the synchronization of multimedia with TTS into text, prosody, the information on synchronization with moving picture, lip-shape, and individual property information in a multimedia information input unit; distributing the information classified in the multimedia information input in a data distributor by each media, based on respective information; converting text distributed in the data distributor by each media into phoneme stream, presuming prosody information and symbolizing the information in a language processor; calculating a value of prosody control parameter other than prosody control parameter included in multimedia information in a prosody processor; adjusting the duration every each phoneme in a synchronization adjustor so that processing result in the prosody processor may be synchronized with a picture signal according to input of the synchronization information; producing the synchronized speech in a signal processor using the prosody

information from the data distributor by each media, the processing result in the synchronization adjustor, and a synthesis unit database; and outputting the picture information distributed by the data distributor by each media onto a screen in a picture output apparatus.

#### BRIEF DESCRIPTION OF THE DRAWINGS

The foregoing and other objects, features, aspects of the present invention will become more apparent from the following detailed description of the present invention when taken in conjunction with the accompanying drawings.

FIG. 1 is a constructional view of a conventional text-to-speech conversion system.

FIG. 2 is a constructional view of a hardware to which the present invention is applied.

FIG. 3 is a constructional view of a text-to-speech conversion system according to the present invention.

#### DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

Now, the present invention will be described in detail by way of the preferred embodiment.

Referring to FIG. 2, a constructional view of hardware to which the present invention is applied is shown. In FIG. 2, the hardware consists of a multimedia data input unit 5, a central



processing unit 6, a synthesis database 7, a digital to analog (D/A) converter 8, and a picture output apparatus 9.

The multimedia data input unit 5 is inputted with data composed of multimedia such as picture and text and outputs this data to the central processing unit 6.

The central processing unit 6 distributes the multimedia data input of the present invention, adjusts synchronization, and performs algorithm based therein to produce synthesized speech.

The synthesis database 7 is a database used in the algorithm for producing the synthesized speech. This synthesis database 7 is stored in a storage device and transmits necessary data to the central processing unit 6.

The digital to analog (D/A) converter 8 converts the synthesized digital data into analog signal and outputs the analog signal.

The picture output apparatus 9 outputs inputted picture information onto a screen.

Table 1 and 2 are algorithms illustrating the state of organized multimedia input information, which consists of text, prosody, the information on synchronization with moving picture, lip-shape, and individual property information.

[Table 1]

Syntax
<pre> TTS_Sequence() {   TTS_Sequence_Start_Code   Prosody_Enable   Video_Enable   Lip_Shape_Enable   Start_Any_Place do{     TTS_Sentence()   }while(next_bits()==TTS_Sentence_Start_Code } </pre>

Here, the TTS\_Sequence\_Start\_Code is a bit string represented with Hexadecimal 'XXXXX' and means a start of TTS sentence.

The TTS\_Sentence\_ID is a 10-bit ID and represents a proper number of each TTS data stream.

The language\_Code represents an object language such as Korean language, English language, German language, Japanese language, French language etc,. to be synthesize.

The prosody\_Enable is a 1-bit flag and has a value of '1' when a prosody data of original sound is included in an organized data.

The Video\_Enable is a 1-bit flag and has a value of '1' when a TTS is interlocked with moving picture.

The Lip\_Shape\_Enable is a 1-bit flag and has a value of '1' when a lip-shape data is included in an organized data.

The Trick\_Mode\_Enable is a 1-bit flag and has a value of '1' when a data is organized to support a trick mode such as stop, restart, forward and backward.

[Table 2]

Syntax
<pre> TTS_Sentence() {   TTS_Sentence_Start_Code   Silence   if(Silence) {     Silence_Duration   }   else {     Gender     Age     if(!Video_Enable) {       Speech_Rate     }     Length_of_Text     TTS_Text     Position_in_Sentence     if(Prosody_Enable) {       Number_of_phonemes       Dur_Enable       F0_Enable       Energy_Enable       for(j=0 ; j&lt;Number_of_phonemes ; j++) {         Symbol_each_phoneme         Dur_each_phoneme         F0_contour_each_phoneme         Energy_contour_each_phoneme       }     }     if(Video_Enable) {       Sentence_Duration       Position_in_Sentence       offset     }     if(Lip_Shape_Enable) {       Number_of_Lip_Event       for(j=0 ; j&lt;Number_of_Lip_Event ; j++) {         Lip_in_Sentence         Lip_Shape       }     }   } } </pre>

Here, the TTS\_Sentence\_Start\_Code is a bit string represented with Hexadecimal 'XXXXX' and means a start of TTS sentence. And the TTS\_Sentence\_Start\_Code is a 10-bit ID and represents a proper number of each TTS data stream.

The TTS\_Sentence\_ID is a 10-bit ID and represents a proper number of each TTS sentence existed in the TTS stream.

The Silence become a '1' when a present input frame of 1-bit flag is silence speech section.

At stage of the Silence\_Duration, a duration time of present silence speech section is represented by milliseconds.

At stage of the Gender, gender is distinguished from a synthesized speech.

At stage of the Age, an age of the synthesized speech distinguished into a baby, youth, middle age and old age.

The Speech\_Rate represents a speech rate of synthesized speech.

At stage of the Length\_of\_Text, a length of input text sentence is represented by byte.

At stage of the TTS\_Text, sentence text having optional

length is represented.

The Dur\_Enable is a 1-bit flag and become a '1' when a duration time information is included in an organized data.

The FO\_Contour\_Enable is a 1-bit flag and become a '1' when a pitch information of each phoneme is included in the organized data.

The Energy\_Contour\_Enable is a 1-bit flag and become a '1' when an energy information of each phoneme is included in the organized data.

At stage of the Number\_of\_Phonemes, the number of phoneme needed to synthesize a sentence are represented.

At stage of the Symbol\_each\_phoneme, symbol such as IPA which is to represent each phoneme is represented.

The Dur\_each\_phoneme represents a duration time of phoneme.

At stage of the FO\_contour\_each\_phoneme, a pitch pattern of the phoneme represented by a pitch value of beginning point, mid point and end point of the phoneme is represented.

At stage of the Energy\_Contour\_each\_phoneme, energy pattern of the phoneme is represented and an energy value of beginning

point, mid point and end point of the phoneme is represented by decibel (dB).

The `Sentence_Duration` represents a total duration time of synthesized speech of the sentence.

The `Position_in_Sentence` represents a position of present frame in the sentence.

At stage of the offset, when the synthesized speech is interlocked with moving picture and a beginning point of the sentence is in the GOP (Group Of Pictures), a delay time consumed from beginning point of GOP to beginning point of the sentence is represented.

The `Number_of_Lip_Event` represents the number of changing point of lip-shape in the sentence.

The `Lip_shape` represents a lip-shape at lip-shape changing point of the sentence.

Text information includes a classification code for a used language and a sentence text. Prosody information includes the number of phoneme in the sentence, phoneme stream information, the duration every each phoneme, pitch pattern of phoneme, energy pattern of phoneme and is used for enhancing the natural of the synthesized speech. The synchronization information of the

moving picture with the synthesized speech can be considered as the dubbing concept and the synchronization could be realized in three ways.

Firstly, there is a method to synchronize between the moving picture and the synthesized speech by the sentence unit by which the duration of the synthesized speech is adjusted using the information about the beginning points of sentences, the durations of sentences, and the delay times of the beginning points of sentences. The beginning points of each sentence indicate locations of scenes from which output of the synthesized speech for each sentence within the moving picture is started. The durations of sentences indicate the number of scenes in which the synthesized speech for each sentence lasts. In addition, the moving picture of MPEG-2 and MPEG-4 picture compression type in which Group of Picture (GOP) concept is used should start at not any scene but a beginning scene within Group of Picture for reproduction. Therefore, the delay time of the beginning point is the information required to synchronize between the Group of Picture and the TTS and indicates a delay time between the beginning scene and a speech beginning point. This method is easy to be realized and can minimize additional effort, but is difficult to accomplish natural synchronization.

Secondly, there is a method by which beginning point information, end point information, and phoneme information are marked every each phoneme within an interval associated with

speech signal in the moving picture and these information is used to produce the synthesized speech. This method has an advantage that degree of accuracy is high since the synchronization between the moving picture and the synthesized speech by the phoneme unit can be attained but a disadvantage that additional effort should be fairly made to detect and record the duration information by the phoneme unit within the speech interval of the moving picture.

Thirdly, there is a method to record the synchronization information based on the beginning point of speech, the end point of speech, lip-shape, and a point of time of lip-shape change. Lip-shape is numeralized to distance (extent of opening) between upper lip and lower lip, distance (extent of width) between left and right and points of lip, and extent of projecting of lip and is defined as a quantized and normalized pattern depended on articulation location and articulation manner of phoneme on the basis of pattern with high discriminative property. This method is a method to raise efficiency of synchronization, while additional effort to produce the information for synchronization can be minimized.

The organized multimedia input information which is applied to the present invention allows an information provider to select and implement optionally among 3 synchronization methods as described above.



In addition, the organized multimedia input information is also used in the process to implement lip animation. Lip animation can be implemented by using phoneme stream prepared from the inputted text in the TTS and the duration every each phoneme, or phoneme stream distributed from the input information and the duration every each phoneme, or by using the information on lip-shape included in the inputted information.

The individual property information allows the user to change gender, age, and speech rate of the synthesized speech.

Gender has male and female, and age is classified into 4, for example, 6-7 years, 18 years, 40 years, and 65 years. The change of speech rate may have 10 steps between 0.7 and 1.6 times of a standard rate. Quality of the synthesized speech can be diversified using these information.

FIG. 3 is a constructional view of the text-to-speech conversion system (TTS) according to the present invention. In FIG. 3, the TTS consists of a multimedia information input unit 10, a data distributor by each media 11, a standardized language processor 12, a prosody processor 13, a synchronization adjustor 14, a signal processor 15, a synthesis unit database 16, and a picture output apparatus 17.

The multimedia input unit 10 is configured as form of Table 1 and 2 and comprises text, prosody information, the information on synchronization with moving picture, the information on lip-

shape. Among these, requisite information is text, other information can be optionally provided by an information provider as optional item for enhancing the individual property and the natural and accomplishing the synchronization with the multimedia, and if needed, can be amended by a TTS user by means of a character input device (keyboard) or a mouse. These information is transmitted to the data distributor by each media 11.

The data distributor by each media 11 receives the multimedia information of which the picture information is transmitted to the picture output apparatus 17, text is transmitted to the language processor 12, and the synchronization information is converted into data structure capable of utilizing in the synchronization adjustor 14 and transmitted to the synchronization adjustor 14. If prosody information is included in the inputted multimedia information, this multimedia information is converted into data structure capable of utilizing in the signal processor 15 and then transmitted to the prosody processor 13 and the synchronization adjustor 14. If individual property information is included in the inputted multimedia information, this multimedia information is converted into data structure capable of utilizing in the synthesis unit database 16 and the prosody processor 13 within the TTS and then transmitted to the synthesis unit database 16 and the prosody processor 13.

The language processor 12 converts text into phoneme stream,

presumes prosody information, symbolizes this information, and then transmits the symbolized information to the prosody processor 13. The symbol of prosody information is presumed from a boundary of the phrase and paragraph, a location of accent in word, a sentence pattern, and so on using the analysis result of syntax.

The prosody processor 13 takes the processing result of the language processor 12 and calculates a value of prosody control parameter other than prosody control parameter included in the multimedia information. Prosody control parameter includes duration pitch contour, energy contour, pause point, and pause length of phoneme. The calculated result is transmitted to the synchronization adjustor 14.

The synchronization adjustor 14 takes the processing result of the prosody processor 13 and adjusts the duration every each phoneme in order to synchronize the result with the picture signal. The adjustment of the duration every each phoneme utilizes the synchronization information transmitted from the data distributor by each media 11. First, lip-shape is assigned to each phoneme depended on articulation location and articulation manner of each phoneme and, on the basis of this, the assigned lip-shape is compared to lip-shape included in the synchronization information and then phoneme stream is divided into small groups by the number of lip-shape recorded in the synchronization information. Also, the duration of phoneme in

the small groups is calculated again using the duration information of lip-shape included in the synchronization information. The adjusted duration information is transmitted to the signal processor 15, included in the processing result of the prosody processor 13.

The signal processor 15 receives the prosody information from the multimedia distributor 11 or the processing result of the synchronization adjustor 14 to produce and output the synthesized speech using the synthesis unit database 16.

The synthesis unit database 16 receives the individual property information from the multimedia distributor 11, selects synthesis units adaptable to gender and age, and then transmits data required for synthesis to the signal processor 15 in response to a request from the signal processor 15.

As can be seen from the description described above, according to the present invention, the individual property of the synthesized speech can be realized and the natural of the synthesized speech can be enhanced by organizing the individual property and prosody information presumed by the analysis of actual speech data, along with text information, as multistage information. Furthermore, a foreign movie can be dubbed in Korean by implementing the synchronization of the synthesized speech with the moving picture by way of the direct use of text information and lip-shape information which is presumed by the

analysis of actual speech data and lip-shape in the moving picture for the production of the synthesized speech. Still furthermore, the present invention is applicable to a variety of field such as communication service, office automation, education and so on by making the synchronization between the picture information and the TTS in the multimedia environment possible.

Although the present invention and its advantages have been described in detail, it should be understood that various changes, substitutions and alterations can be made herein without departing from the spirit and scope of the invention as defined by the appended claims.

It is therefore intended by the appended claims to cover any and all such applications, modifications, and embodiments within the scope of the present invention.

WHAT IS CLAIMED IS:

1. A text-to-speech conversion system (TTS) for interlocking with multimedia comprising:

a multimedia information input unit for organizing text, prosody, the information on synchronization with moving picture, lip-shape, and the information such as individual property;

a data distributor by media for distributing the information of the multimedia information input unit into the information by each media;

a language processor for converting the text distributed by the data distributor by each media into phoneme stream, presuming prosody information and symbolizing the information;

a prosody processor for calculating a value of prosody control parameter from the symbolized prosody information using a rule and a table;

a synchronization adjustor for adjusting the duration of the phoneme using the synchronization information distributed by the data distributor by each media;

a signal processor for producing a synthesized speech using the prosody control parameter and data in a synthesis unit database; and

a picture output apparatus for outputting the picture information distributed by the data distributor by each media onto a screen.

2. A method for organizing input data of a text-to-speech

conversion system (TTS) for interlocking with multimedia comprising the steps of:

classifying multimedia input information organized for enhancing the natural of synthesized speech and implementing the synchronization of multimedia with TTS into text, prosody, the information on synchronization with moving picture, lip-shape, and individual property information in a multimedia information input unit;

distributing the information classified in the multimedia information input in a data distributor by each media, based on respective information;

converting text distributed in the data distributor by each media into phoneme stream, presuming prosody information and symbolizing the information in a language processor; calculating a value of prosody control parameter other than prosody control parameter included in multimedia information in a prosody processor;

adjusting the duration every each phoneme in a synchronization adjustor so that processing result in the prosody processor may be synchronized with a picture signal according to input of the synchronization information;

producing the synchronized speech in a signal processor using the prosody information from the data distributor by each media, the processing result in the synchronization adjustor, and a synthesis unit database; and

outputting the picture information distributed by the data distributor by each media onto a screen in a picture output

apparatus.

3. The method according to claim 2, wherein said organized multimedia information is comprised of text information, prosody information, information synchronized with moving picture, lip-shape and individuality information.

4. The method according to claim 3, wherein said prosody information is comprised of the number of phoneme, phoneme stream information, duration time of each phoneme, pitch pattern of the phoneme and energy pattern of the phoneme.

5. The method according to claim 4, wherein said duration of the phoneme is indicative of a value of pitch at beginning point, middle point, and end point within the phoneme.

6. The method according to claim 4, wherein said energy pattern of the phoneme is indicative of a value of energy in decibel at beginning point, mid point and end point within phoneme.

7. The method according to claim 2, wherein said synchronization information is comprised of text, lip-shape, location information with moving picture, and the duration information.

8. The method according to claim 2, wherein said synchronization information is composed of a beginning point, duration and delay time information of starting point, and duration of each phoneme



is controlled by said synchronization information.

9. The method according to claim 2, wherein said synchronization information is composed of a duration of the beginning point of a sentence and a duration information of starting point, and duration of each phoneme is controlled by forecast lip-shape considered an articulation manner of the phoneme and articulation control,

lip-shape within the synchronization and duration information composed of said synchronization information.

10. The method according to claim 2, wherein said synchronized speech is produced by an information of beginning point and end point of each phoneme related with speech signal and an information of phoneme.

11. The method according to claim 2, wherein said synchronized speech is produced by a numeralization of distance(extent of opening) between upper lip and low lip, distance(extent of width) between left and right end points of lip, and extent of projecting of lip and the lip-shape quantized and normalized pattern depended on articulation location and articulation manner of the phoneme on the basis of pattern with high discriminative property.

12. The method according to claim 2, wherein said transmission method of multimedia information comprising the steps of:

converting a prosody information existed in the multimedia information into a data structure capable of utilizing in the signal processor;

transmitting the converted prosody information to the prosody and the synchronization adjustor;

converting the prosody information outputed from the prosody and the synchronization adjustor to a data structure capable of utilizing in the synthesis unit database and the prosody processor within the TTS if the prosody information is included in said multimedia input information;

transmitting then to the synthesis unit database and the prosody processor if the individual property information is included in said multimedia input information.

0902071-09090

## ABSTRACT

This invention relates to a text-to-speech conversion system (TTS) for interlocking with multimedia and a method for organizing input data of the same. A conventional TTS is in situation of the only for the synthesis of speech from the inputted text. In addition, by a prior organization, since it is impossible to presume from only the text the information required when moving picture is to be dubbed by use of TTS or when the natural interlock between the synthesized speech and multimedia such as animation is to be implemented, there is no method to realize these function. Furthermore, there is also no result of the studies on use of additional data for enhancement of the natural in the synthesized speech and organization of these data. Therefore, an object of the present invention is to provide a text-to-speech conversion system (TTS) for interlocking multimedia and a method for organizing input data of the same for enhancing the natural of synthesized speech and accomplishing the synchronization of multimedia with TTS by defining additional prosody information, the information required to interlock TTS with multimedia, and interface between these information and TTS for use in the production of the synthesized speech. According to the present invention, a foreign movie can be dubbed in Korean by implementing the synchronization of the synthesized speech with the moving picture by way of the direct use of text information and lip-shape information which is presumed by the

[illegible]

Attorney Docket # 4577-50

Patent

**IN THE UNITED STATES PATENT AND TRADEMARK OFFICE**

In re Application of

Jung Chul LEE et al.

Serial No.: Not Yet Assigned

Filed: Concurrently herewith

For: A Text-To-Speech Conversion System  
For Interlocking With Multimedia And A  
Method For Organizing Input Data Of  
The Same

**PRELIMINARY AMENDMENT**

Assistant Commissioner for Patents  
Washington, D.C. 20231

S I R:

Prior to examination of the above-identified application please amend the  
application as follows:

In the Specification:

Page 8, delete the title "[Table 1]" and the table itself and substitute therefor the title --Table 1-- and the following table:

Syntax
TTS_Sequence( ) { TTS_Sequence_Start_Code TTS_Sentence_ID Language_Code Prosody_Enable Video_Enable Lip_Shape_Enable Trick_Mode_Enable do{ TTS_Sentence ( ) }while (next_bits( ) == TTS_Sentence_Start_Code }

Page 9, delete the title "[Table 2]" and the table itself and substitute therefor the title --Table 2-- and the following table:

## Syntax

```

TTS_Sentence( ) {
TTS_Sentence_Start_Code
TTS_Sentence_ID
Silence
if (Silence) {
    Silence_Duration
}
else {
    Gender
    Age
    if(!Video_Enable) {
        Speech_Rate
    }
    Length_of_Text
    TTS_Text()
    if(Prosody_Enable) {
        Dur_Enable
        FO_Contour_Enable
        Energy_Contour_Enable
        Number_of_Phonemes
        for(j=0 ; j<Number_of_phonemes ; j++) {
            Symbol_each_phoneme
            if(Dur_Enable) {
                Dur_each_phoneme
            }
            if(FO_Contour_Enable {
                FO_contour_each_phoneme
            }
            if(Energy_Contour_Enable) {
                Energy_contour_each_phoneme
            }
        }
    }
    if(Video_Enable) {
        Sentence_Duration
        Position_in_Sentence
        offset
    }
    if(Lip_Shape_Enable) {
        Number_of_Lip_Event
        for(j=0 ; j<Number_of_Lip_Event ; j++) {
            Lip_in_Sentence
            Lip_shape
        }
    }
}
}
}

```

In the Claims:

Please cancel claims 1-12, without prejudice.

Please add the following new claims:

--13. A text-to-speech conversion system for interlocking with multimedia comprising:

a multimedia information input unit for organizing text, prosody information, information on synchronization with a moving picture, lip-shape information, picture information, and individual property information;

a data distributor by each media for distributing the information of said multimedia information input unit into information for each media;

a language processor for converting the text distributed by said data distributor by each media into a phoneme stream, presuming prosody information and symbolizing the presumed prosody information;

a prosody processor for calculating a prosody control parameter value from the symbolized prosody information;

a synchronization adjustor for adjusting a duration of each phoneme using the synchronization information distributed by said data distributor by each media;

a synthesis unit database for receiving the individual property information from said data distributor by each media, selecting synthesis units adaptable to gender and age, and outputting data required for synthesis;



a signal processor for producing a synthesized speech using the prosody control parameter and the data output from said synthesis unit database; and

a picture output apparatus for outputting the picture information distributed by said data distributor by each media on to a screen.

14. A method for organizing input data of a text-to-speech conversion system for interlocking with multimedia, said method comprising the steps of:

(a) classifying multimedia input information organized for enhancing natural synthesized speech and implementing synchronization of multimedia with text-to-speech into text, prosody information, information on synchronization with a moving picture, lip-shaped information, picture information, and individual property information using a multimedia information input unit;

(b) distributing using a data distributor by each media the multimedia input information classified in the multimedia information input unit based on respective information;

(c) converting the text distributed by the data distributor by each media into a phoneme stream, presuming prosody information and symbolizing the presumed prosody information using a language processor;

(d) calculating a prosody control parameter value other than a prosody control parameter included in the multimedia input information using a prosody processor;

(e) adjusting a duration of each phoneme using a synchronization adjustor so as to synchronize a processing result of the prosody processor with a picture signal according to the synchronization information distributed by the data distributor by each media;

(f) selecting synthesis units adaptable to gender and age based on the individual property information from the data distributor by each media using a synthesis unit database and outputting data required for synthesis;

(g) producing synthesized speech using a signal processor based on the prosody information distributed by the data distributor by each media, a processing result of the synchronization adjustor, and the data from the synthesis unit database; and

(h) outputting the picture information distributed by the data distributor by each media onto a screen using a picture output unit.

15. The method in accordance with claim 14, wherein the organized multimedia information comprises text information, prosody information, information on synchronization with a moving picture, lip-shaped information, and individual property information.

16. The method in accordance with claim 15, wherein the prosody information comprises a number of phoneme, phoneme stream information, duration of each phoneme, pitch pattern of the phoneme, and energy pattern of the phoneme.

17. The method in accordance with claim 16, wherein the duration time of the phoneme is indicative of a value of pitch at a beginning point, a mid point, and an end point within the phoneme.

18. The method in accordance with claim 17, wherein the energy pattern of the phoneme is indicative of a value of energy in decibels at the beginning point, the mid point, and the end point within the phoneme.

19. The method in accordance with claim 15, wherein the synchronization information comprises text, lip-shape, location information with a moving picture, and duration information.

20. The method in accordance with claim 15, wherein the synchronization information comprises a beginning point, duration and delay time information of a starting point, and duration of each phoneme is controlled by the synchronization information.

21. The method in accordance with claim 15, wherein the synchronization information is composed of a duration of a beginning point of a sentence, a duration information of a starting point, and duration of each phoneme is controlled by forecast lip-shape considered an articulation manner of the phoneme and articulation control of lip-shape within the synchronization and duration information of the synchronization information.

22. The method in accordance with claim 15, wherein the synthesized speech is produced based on beginning point information, end point information, and phoneme information for each phoneme within an interval associated with a speech signal.

23. The method in accordance with claim 15, wherein the synthesized speech is produced based on a distance of an opening between an upper lip and a lower lip, a distance between end points of the lips, and an extent of projection of a lip, and a lip-shape quantized and normalized pattern is defined depending on articulation location and articulation manner of the phoneme on a basis of pattern with discriminative property.

24. The method in accordance with claim 15, wherein if the multimedia input information comprises prosody information, further comprising the steps of:

(i) converting the prosody information into a data structure recognizable by the signal processor; and

(j) transmitting the converted prosody information the prosody processor and the synchronization adjustor.

25. The method in accordance with claim 15, wherein if the multimedia input information includes individual property information, further comprising the steps of:

(k) converting the individual property information into a data structure recognizable by the synthesis unit database and the prosody processor within the text-to-speech;

(1) transmitting the converted individual property information to the synthesis unit database and the prosody processor.--

**REMARKS**

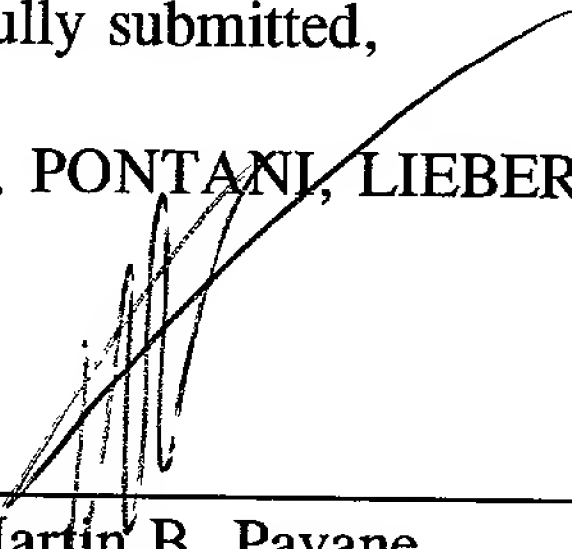
This preliminary amendment is presented to place the claims in better condition for examination. No new matter has been added. Early examination and favorable consideration of the above-identified application is earnestly solicited.

Any additional fees or charges required at this time in connection with the application may be charged to our Patent and Trademark Office Deposit Account No. 03-2412.

Respectfully submitted,

COHEN, PONTANI, LIEBERMAN & PAVANE

By

  
\_\_\_\_\_  
Martin B. Pavane  
Reg. No. 28,337  
551 Fifth Avenue, Suite 1210  
New York, N.Y. 10176  
(212) 687-2770

February 9, 1998

FIG. 1

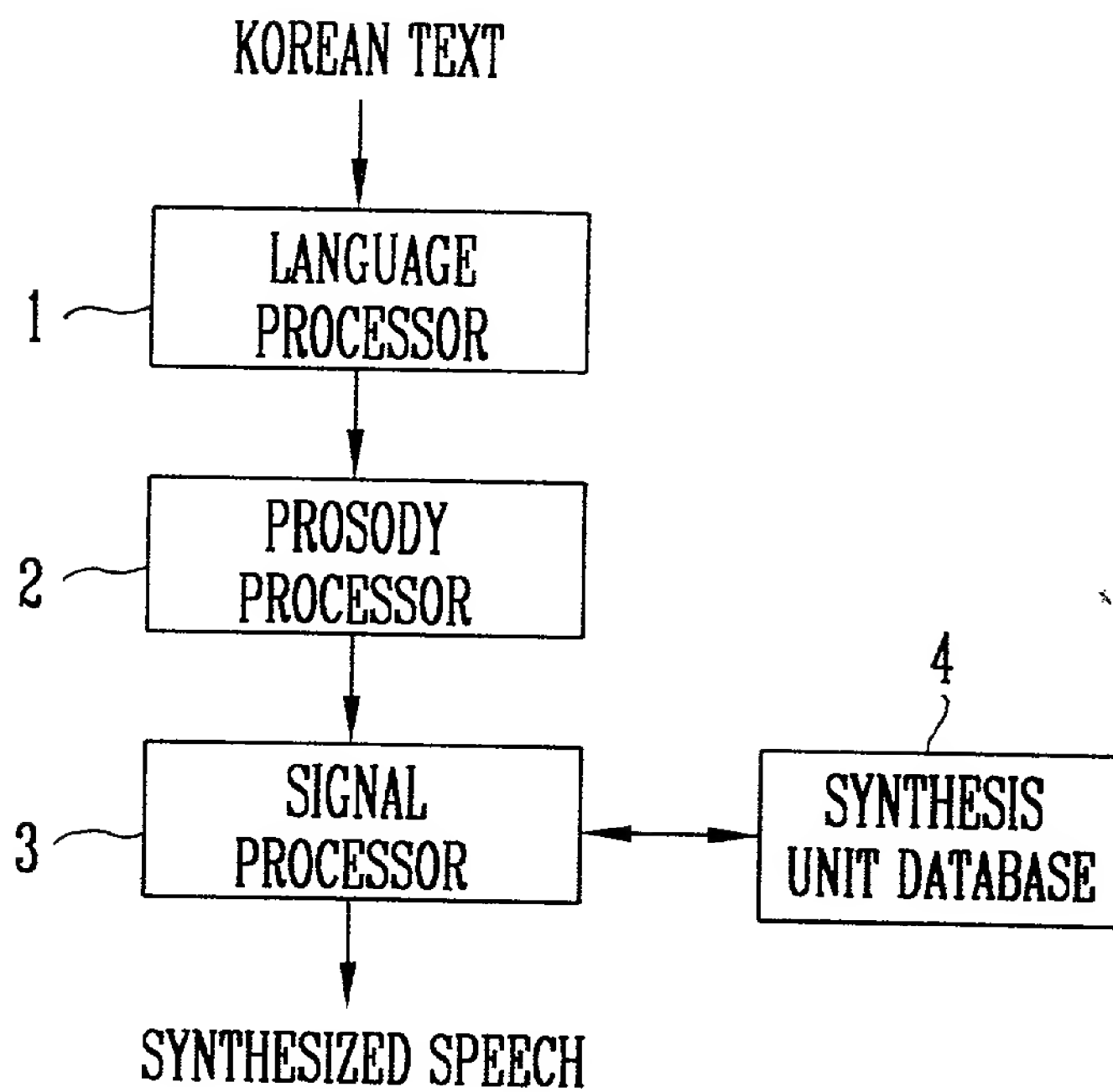


FIG. 2

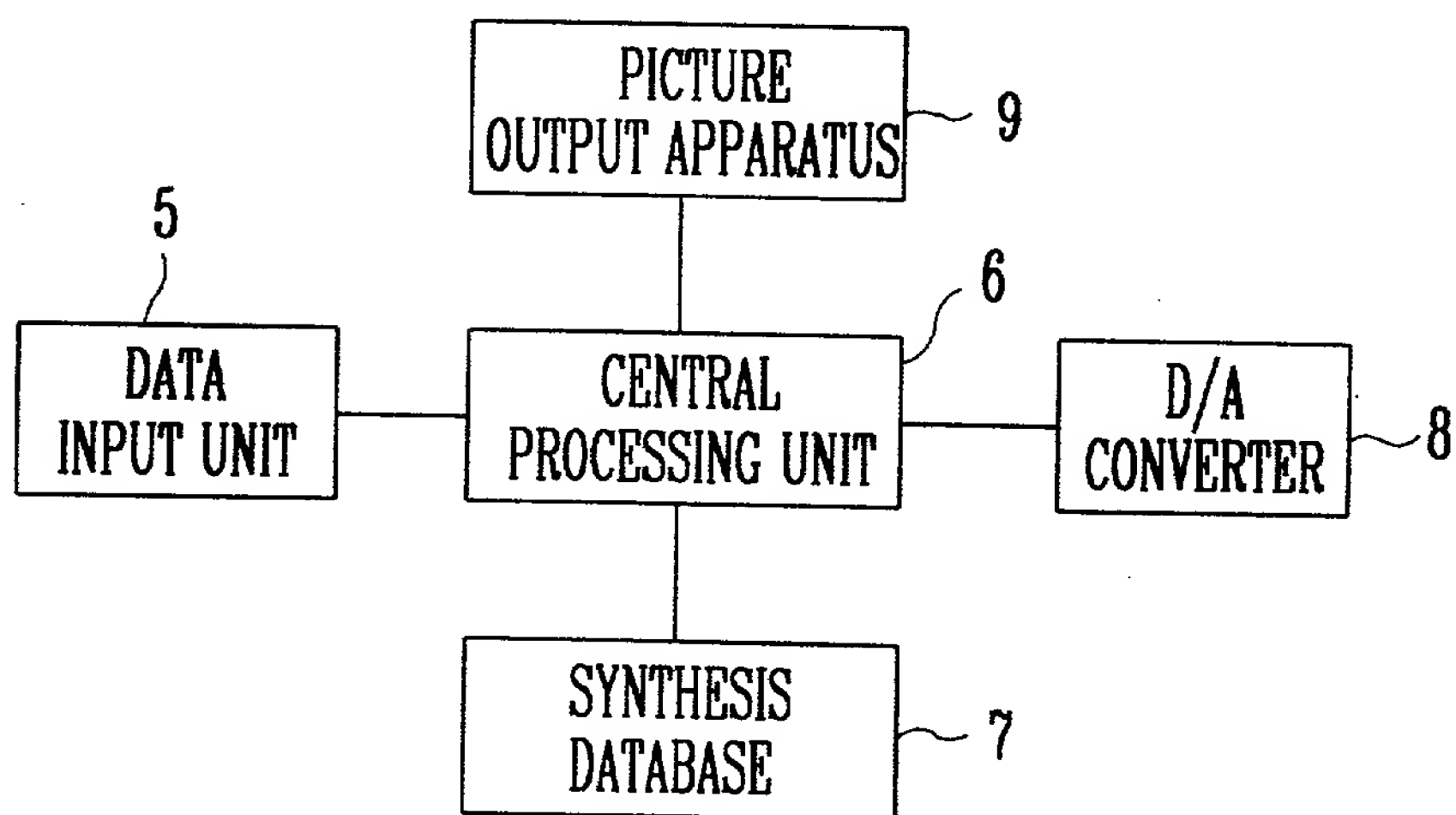
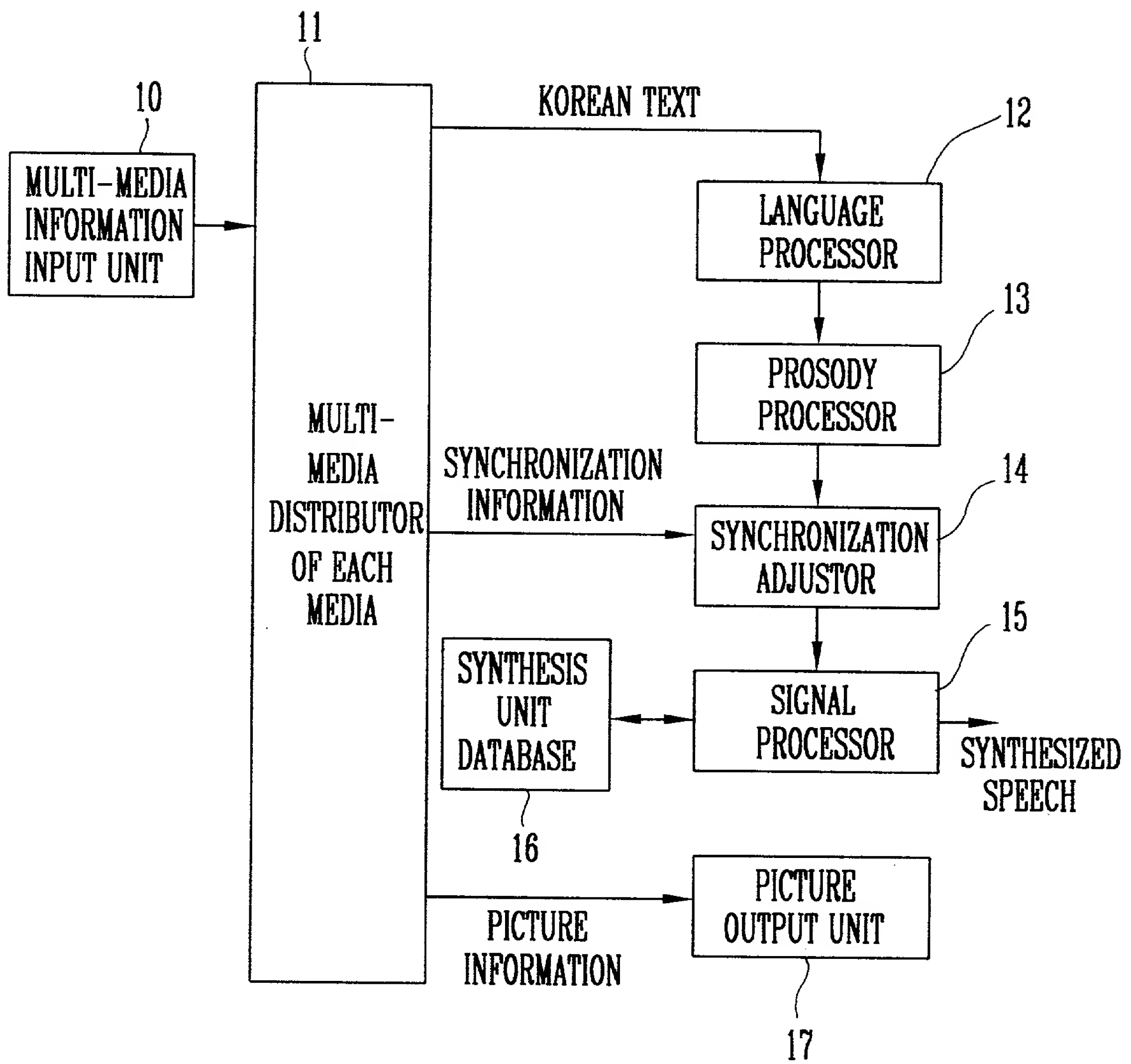


FIG. 3



VERIFIED STATEMENT (DECLARATION) CLAIMING SMALL ENTITY  
STATUS (37 CFR 1.9(f) and 1.27(b)) - SMALL BUSINESS CONCERN

Applicant or Patentee: \_\_\_\_\_  
Attorney's Docket No.: \_\_\_\_\_  
Serial or Patent No.: \_\_\_\_\_  
Filed or Issued: \_\_\_\_\_

Title: A TEXT-TO-SPEECH CONVERSION SYSTEM FOR INTERLOCKING WITH  
MULTIMEDIA AND A METHOD FOR ORGANIZING INPUT DATA OF THE SAME

I hereby declare that I am

- ☐ the owner of the small business concern identified below:  
☐ an official of the small business concern empowered to act on behalf of the concern identified below:

NAME OF SMALL BUSINESS CONCERN: \_\_\_\_\_

ADDRESS OF CONCERN: \_\_\_\_\_

I hereby declare that the above identified small business concern qualifies as a small business concern as defined in 13 CFR 121.12, and reproduced in 37 CFR 1.9(d), for purposes of paying reduced fees to the United States Patent and Trademark Office in that the number of employees of the concern, including those of its affiliates, does not exceed 500 persons. For purposes of this statement, (1) the number of employees of the business concern is the average over the previous fiscal year of the concern of the persons employed on a full-time, part-time or temporary basis during each of the pay periods of the fiscal year, and (2) concerns are affiliates of each other when either, directly or indirectly, one concern controls or has the power to control the other, or a third party or parties controls or has the power to control both.

I hereby declare that rights under contract or law have been conveyed to and remain with the small business concern identified above with regard to the invention described in:

- ☐ the specification filed herewith with title listed above.  
☐ the application identified above.  
☐ the patent identified above.

If the rights held by the above identified small business concern are not exclusive, each individual, concern or organization having rights to the invention must file separate verified statements averring to their status as small entities, and no rights to the invention are held by any person, other than the inventor, who would not qualify as an independent inventor under 37 CFR 1.9(c) if that person made the invention or by any concern which would not qualify as a small business concern under 37 CFR 1.9(d), or a nonprofit organization under 37 CFR 1.9(e).

Each person, concern or organization having any fights in the invention is listed below:

- ☐ no such person, concern, or organization exists.  
☐ each such person, concern or organization listed below

FULL NAME: ELECTRONICS AND TELECOMMUNICATIONS RESEARCH INSTITUTE

ADDRESS: 161 Kajong-Dong, Yusong-Gu, Daejon-Shi, Korea

☐ INDIVIDUAL ☐ SMALL BUSINESS CONCERN ☐ NONPROFIT ORGANIZATION

Separate verified statements are required from each named person, concern or organization having rights to the invention averring to their status as small entities. (37 CFR 1.27)

I acknowledge the duty to file, in this application or patent, notification of any change in status resulting in loss of entitlement to small entity status prior to paying, or at the time of paying, the earliest of the issue fee or any maintenance fee due after the date on which status as a small entity is no longer appropriate. (37 CFR 1.28(b))

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under section 1001 of Title 18 of the United States Code, and that such willful false statements may jeopardize the validity of the application, any patent issuing thereon, or any patent to which this verified statement is directed.

NAME OF PERSON SIGNING Keun Jang Song

TITLE OF PERSON OTHER THAN OWNER Head of Intellectual Property Section

ADDRESS OF PERSON SIGNING 161 Kajong-Dong, Yusong-Gu, Daejon-Shi, Korea

SIGNATURE KS Song DATE 20/10/ 1997



**COMBINED DECLARATION FOR PATENT APPLICATION AND POWER OF ATTORNEY**  
Includes Reference to PCT International Applications

Attorney's Docket No.

As a below named inventor, I hereby declare that:

My residence, post office address and citizenship are as stated below next to my name.

I believe I am the original, first and sole inventor (if only one name is listed below) or an original, first and joint inventor (if plural names are listed below) of the subject matter which is claimed and for which a patent is sought on the invention entitled:

**A TEXT-TO-SPEECH CONVERSION SYSTEM FOR INTERLOCKING WITH MULTIMEDIA  
AND A METHOD FOR ORGANIZING INPUT DATA OF THE SAME**

the specification of which (check only one item below)

☐ is attached hereto

☐ was filed as United States application

Serial No. \_\_\_\_\_

on \_\_\_\_\_

and was amended

on \_\_\_\_\_ (if applicable).

☐ was filed as PCT international application

Number \_\_\_\_\_

on \_\_\_\_\_

and was amended under PCT Article 19

on \_\_\_\_\_ (if applicable).

I hereby state that I have reviewed and understand the contents of the above-identified specification, including the claims, as amended by any amendment referred to above.

I acknowledge the duty to disclose information which is material to the patentability of the application in accordance with Title 37, Code of Federal Regulations, §.56(a).


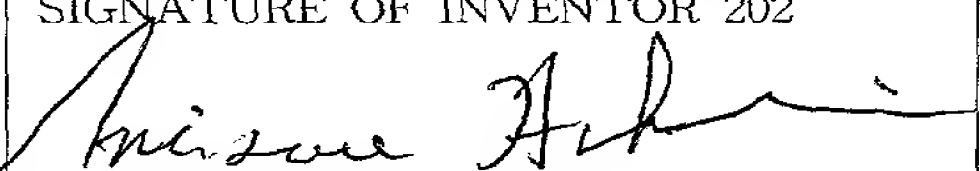
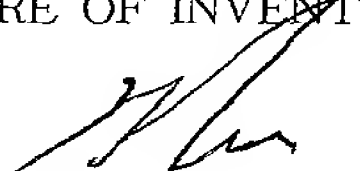
I hereby claim foreign priority benefits under Title 35, United States Code, §19 of any foreign application(s) for patent or inventor's certificate or of any PCT international application(s) designating at least one country other than the United States of America listed below and have also identified below any foreign application(s) for patent or inventor's certificate or any PCT international application(s) designating at least one country other than the United States of America filed by me on the same subject matter having a filing date before that of the application(s) of which priority is claimed.

**PRIOR FOREIGN/PCT APPLICATIONS AND ANY PRIORITY CLAIMS UNDER 35 U.S.C. 119:**

Country (if PCT, indicate "PCT")	Application Number	Date of Filing (day, month, year)	Priority Claimed Under 35 U.S.C. 119	
Korea	97-17615	08/05/1997	<input checked="" type="checkbox"/> YES	<input type="checkbox"/> NO
			<input type="checkbox"/> YES	<input type="checkbox"/> NO
			<input type="checkbox"/> YES	<input type="checkbox"/> NO
			<input type="checkbox"/> YES	<input type="checkbox"/> NO
			<input type="checkbox"/> YES	<input type="checkbox"/> NO

<b>COMBINED DECLARATION FOR PATENT APPLICATION AND POWER OF ATTORNEY (Continued)</b> Includes Reference to PCT International Applications				<b>Attorney's Docket No.</b>	
I hereby claim the benefit under Title 35, United States Code, § 120 of any United States application(s) or PCT international application(s) designating the United States of America that is/are listed below and, insofar as the subject matter of each of the claims of this application is not disclosed in that/those prior application(s) in the manner provided by the first paragraph of Title 35, United States Code, § 112, I acknowledge the duty to disclose material information as defined in Title 37, Code of Federal Regulations, § 1.56(a) which occurred between the filing date of the prior application(s) and the national or PCT international filing date of this application:					
<b>PRIOR U.S. APPLICATIONS OR PCT INTERNATIONAL APPLICATIONS DESIGNATING THE U.S. FOR BENEFIT UNDER 35 U.S.C. 120:</b>					
U.S. APPLICATIONS			STATUS (check one)		
U.S. APPLICATION NUMBER	U.S. FILING DATE	PATENTED	PENDING	ABANDONED	
PCT APPLICATIONS DESIGNATING THE U.S.					
PCT APPLICATION NO.	PCT FILING DATE	U.S. SERIAL NUMBERS ASSIGNED (if any)			
<b>POWER OF ATTORNEY:</b> As a named inventor, I hereby appoint the following attorney(s) and/or agent(s) to prosecute this application and transact all business in the Patent and Trademark Office connected therewith (List name and registration number)  MYRON COHEN, Reg. No. 17,358; THOMAS C. PONTANI, Reg. No. 29,763; LANCE J. LIEBERMAN, Reg. No. 28,437; MARTIN B. PAVANE, Reg. No. 28,337; MICHAEL C. STUART, Reg. No. 35,698; JAMES J. DeCARLO, Reg. No. 36,120; CAROL E. ROZEK, Reg. No. 36,993; EDWARD M. WEISZ, Reg. No. 37,257; KLAUS P. STOFFEL, Reg. No. 31,668; CHI K. ENG, Reg. No. 38,870; EDWARD ETKIN, Reg. No. 37,824; CHERYL COHEN, Reg. No. 40,361; and JULIA S. KIM, Reg. No. 36,567.					
Send correspondence to:  Martin B. Pavane, Esq. Reg. No. 28,337 Cohen, Pontani, Lieberman & Pavane 551 Fifth Avenue, Suite 1210 New York, New York 10176			Direct Telephone calls to: (name and telephone number)  Martin B. Pavane, Esq. (212) 687-2770		
201	FULL NAME OF INVENTOR	FAMILY NAME <b>Lee</b>	FIRST GIVEN NAME <b>Jung</b>	SECOND GIVEN NAME <b>Chul</b>	
	RESIDENCE & CITIZENSHIP	CITY <b>Daejon-Shi</b>	STATE OR FOREIGN COUNTRY <b>Korea</b>	COUNTRY OF CITIZENSHIP <b>Korea</b>	
	POST OFFICE ADDRESS	POST OFFICE ADDRESS <b>Sambu Apt. 36-102, Taepyung-Dong, Choong-Gu</b>	CITY <b>Daejon-Shi</b>	STATE & ZIP CODE/COUNTRY <b>301-150 / Korea</b>	
202	FULL NAME OF INVENTOR	FAMILY NAME <b>Hahn</b>	FIRST GIVEN NAME <b>Min</b>	SECOND GIVEN NAME <b>Soo</b>	
	RESIDENCE & CITIZENSHIP	CITY <b>Daejon-Shi</b>	STATE OR FOREIGN COUNTRY <b>Korea</b>	COUNTRY OF CITIZENSHIP <b>Korea</b>	
	POST OFFICE ADDRESS	POST OFFICE ADDRESS <b>Hanwool Apt. 106-1004, Shinsung-Dong, Yusong-Gu</b>	CITY <b>Daejon-Shi</b>	STATE & ZIP CODE/COUNTRY <b>305-345 / Korea</b>	

Equivalent to PTO 139(REV.10 83)

Combined Declaration for Patent Application and Power of Attorney (Continued) (Includes Reference to PCT International Applications)				Attorney's Docket No.
203	FULL NAME OF INVENTOR	FAMILY NAME Lee	FIRST GIVEN NAME Hang	SECOND GIVEN NAME Seop
	RESIDENCE & CITIZENSHIP	CITY Daejon-Shi	STATE OR FOREIGN COUNTRY Korea	COUNTRY OF CITIZENSHIP Korea
	POST OFFICE ADDRESS	POST OFFICE ADDRESS Chowan Apt. 106-1509, Mannyeon-Dong, Seo-Gu	CITY Daejon-Shi	STATE & ZIP CODE/COUNTRY 302-150 / Korea
204	FULL NAME OF INVENTOR	FAMILY NAME	FIRST GIVEN NAME	SECOND GIVEN NAME
	RESIDENCE & CITIZENSHIP	CITY	STATE OR FOREIGN COUNTRY	COUNTRY OF CITIZENSHIP
	POST OFFICE ADDRESS	POST OFFICE ADDRESS	CITY	STATE & ZIP CODE/COUNTRY
205	FULL NAME OF INVENTOR	FAMILY NAME	FIRST GIVEN NAME	SECOND GIVEN NAME
	RESIDENCE & CITIZENSHIP	CITY	STATE OR FOREIGN COUNTRY	COUNTRY OF CITIZENSHIP
	POST OFFICE ADDRESS	POST OFFICE ADDRESS	CITY	STATE & ZIP CODE/COUNTRY
206	FULL NAME OF INVENTOR	FAMILY NAME	FIRST GIVEN NAME	SECOND GIVEN NAME
	RESIDENCE & CITIZENSHIP	CITY	STATE OR FOREIGN COUNTRY	COUNTRY OF CITIZENSHIP
	POST OFFICE ADDRESS	POST OFFICE ADDRESS	CITY	STATE & ZIP CODE/COUNTRY
<p>I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment or both, under §1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issuing thereon.</p>				
SIGNATURE OF INVENTOR 201 		SIGNATURE OF INVENTOR 202 		SIGNATURE OF INVENTOR 203 
DATE 20/10/1997		DATE 20/10/1997		DATE 20/10/1997
SIGNATURE OF INVENTOR 204		SIGNATURE OF INVENTOR 205		SIGNATURE OF INVENTOR 206
DATE		DATE		DATE